Scandinavian Journal of Forest Research
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713711862

A model-assisted *k*-nearest neighbour approach to remove extrapolation bias

Steen Magnussen[a]; Erkki Tomppo[b]; Ronald E. McRoberts[c]
[a] Natural Resources Canada, Canadian Forest Service, Victoria, BC, Canada [b] Finnish Forest Research Institute, Vantaa, Finland [c] USDA Forest Service, Northern Research Station, Minnesota, MN, USA

First published on: 10 March 2010

To cite this Article Magnussen, Steen , Tomppo, Erkki and McRoberts, Ronald E.(2010) 'A model-assisted *k*-nearest neighbour approach to remove extrapolation bias', Scandinavian Journal of Forest Research, 25: 2, 174 — 184, First published on: 10 March 2010 (iFirst)
To link to this Article: DOI: 10.1080/02827581003667348
URL: http://dx.doi.org/10.1080/02827581003667348

Scandinavian Journal of Forest Research, 2010; 25: 174–184

Taylor & Francis
Taylor & Francis Group

ORIGINAL ARTICLE

A model-assisted *k*-nearest neighbour approach to remove extrapolation bias

STEEN MAGNUSSEN[1], ERKKI TOMPPO[2] & RONALD E. MCROBERTS[3]

[1]*Natural Resources Canada, Canadian Forest Service, 506 West Burnside Rd., Victoria BC V8Z 1M5, Canada,* [2]*Finnish Forest Research Institute, Vantaa, FI-01301 Finland,* [3]*USDA Forest Service, Northern Research Station, St Paul, Minnesota, MN 55108, USA*

Abstract
In applications of the *k*-nearest neighbour technique (kNN) with real-valued attributes of interest (**Y**) the predictions are biased for units with ancillary values of **X** with poor or no representation in a sample of *n* units. In this article a model-assisted calibration is proposed that reduces unit-level extrapolation bias. The bias is estimated as the difference in model-based predictions of **Y** given the **X**-values of the true *k* nearest units and the *k* selected reference units. Calibrated kNN predictions are then obtained by adding this difference to the original kNN prediction. The relationship is modelled between **Y** and **X** with decorrelated **X**-variables, variables scaled to the interval [0,1] and Bernstein basis functions to capture changes in **Y** as a function of changes in **X**. Three examples with actual forest inventory data from Italy, the USA and Finland demonstrated that calibrated kNN predictions were, on average, closer to their true values than non-calibrated predictions. Calibrated predictions had a range much closer to the actual range of **Y** than non-calibrated predictions.

Keywords: *Bernstein basis functions, extrapolation bias, multivariate calibration, non-parametric prediction.*

Introduction

The *k*-nearest neighbours (kNN) technique is an appealing non-parametric approach to either univariate or multivariate prediction of a desired attribute (**Y**) based on the similarity in the ancillary variable space (**X**) between a (target) unit for which a prediction is desired, and a set of reference units for which an observation of **Y** is available (Alt, 2001). The kNN prediction is a weighted linear sum of the *k* reference units that are nearest in terms of **X**-values to the target unit. Its ease and flexibility have made it popular in forestry applications (Franco-Lopez et al., 2001; Meng et al., 2007; LeMay et al., 2008; Tomppo et al., 2008). However, the technique is inherently biased, because no prediction may be smaller (larger) than the smallest (largest) observed **Y**-value.

Unlike model-based statistical prediction methods, the kNN technique is particularly vulnerable to poor performance when the **X**-value of a target unit has no close neighbours in the reference set. With the kNN technique, a prediction for a unit with ancillary values outside the **X**-domain defined by the reference units will be identical to the prediction made for the unit with the closest **X**-value inside the domain of the reference units (Stage & Crookston, 2007; Fehrmann et al., 2008; McRoberts, 2009). The net effect is a tendency to underestimate large **Y**-values and overestimate small **Y**-values, not unlike a regression to the mean effect (Krause & Pinheiro, 2007). Gaps between neighbouring **X**-values in a sample of *n* reference units will be larger than in the population (Stage & Crookston, 2007) which, everything else being equal, introduces another source of bias.

Experience has shown kNN estimates of a univariate total to be nearly unbiased, at least when the sample is relatively large and representative of the population (Katila, 2006; Magnussen et al., 2009). The unit-level bias is therefore mainly a concern in unit-level applications, e.g. in databases (Stage &

Correspondence: S. Magnussen, Natural Resources Canada, Canadian Forest Service, 506 West Burnside Rd., Victoria BC V8Z 1M5, Canada. E-mail: steen.magnussen@nrcan.gc.ca

(Received 7 October 2009; accepted 18 January 2010)

ISSN 0282-7581 print/ISSN 1651-1891 online
DOI: 10.1080/02827581003667348

Crookston, 2007; Tomppo et al., 2008) and spatial map displays (LeMay et al., 2008).

One way to address the extrapolation problem is to lower the weights given to variables with poor coverage (McRoberts, 2009) or outright drop them (Stage & Crookston, 2007). However, with many ancillary variables the "curse of dimensionality" (Scott, 1992, p. 195) makes it a daunting task to find a kNN weighting scheme that achieves a desired outcome (Tomppo & Halme, 2004). Using a smaller number ($k^{\star}$) of reference units than the k that minimizes the root mean-squared difference can, to some extent, reduce the extrapolation bias, but the gain is at the expense of a less efficient estimator (Stage & Crookston, 2007).

In this study a model-based calibration of unit-level kNN predictions is proposed that achieves a correction of the relationship between actual **Y**-values and the kNN predictions towards a 1:1 line with an intercept of zero. The calibration requires model-based predictions of the effect of using the k-nearest reference units instead of the actual k-nearest units in the population for making a prediction. To this end, a simple working linear model for the relationship between **X** and **Y** is entertained. The model should provide robust and realistic estimates of the effect on **Y** of a change in **X**. A search for the best model is not needed as it would negate any attraction of kNN. Instead, the flexible Bernstein basis functions are used to capture the relationship between **Y** and **X** (Lorentz, 1953). The calibration procedure is demonstrated in simulated cluster and simple random sampling (without replacement) in artificial populations composed of actual unit-level (plot) forest inventory data (**Y**) and Landsat ETM+ derived ancillary variables (**X**).

Materials and methods

Unit-level k-nearest neighbour estimator

Let U be a population composed of N units. We are interested in a set of q attributes **Y** for the purpose of estimating unit-level values of y_i, $i=1, \ldots, N$ used to estimate the population total (T_y) and in small-area estimation. **Y**-values are only known for units in a sample (**s**) of size n. The n sample units are referred to as reference units and it is assumed that they arise from probability sampling (Hansen et al., 1983). A set of p ancillary variables **X** carrying information about **Y** is known for every unit in the population. A unit for which a kNN prediction is made is called a target unit.

A general unit-level kNN estimator of **Y** for the ith population unit (Haara et al., 1997) is

$$\tilde{\mathrm{y}}_i = \sum_{i \sim k^{\mathrm{j}}} \mathbf{w}_j y_j, \quad j \in \mathbf{s}, \;\; i = 1, \ldots, N, \tag{1}$$

where w_j is a $q \times q$ matrix of weights and $i \sim k^{\mathrm{j}}$ means that the summation is taken over the k (reference) units $\mathbf{x}_j$ nearest to $\mathbf{x}_i$ with respect to some distance metric. For the sake of demonstration, $\tilde{\mathrm{y}}_i$ is computed as the arithmetic mean of the **Y**-values of the k selected reference units, i.e. $\mathbf{w}_j$ in eq. (1) is a diagonal matrix with elements k^{-1} along the diagonal and 0 elsewhere. Euclidean distances between unit-level vectors of **X**-values are used to identify the k-nearest neighbours. The kNN estimator of a population total becomes

$$\tilde{\mathbf{T}}_y = \sum_{i=1}^{N} \tilde{\mathbf{y}}_i. \tag{2}$$

A unit average is estimated by dividing $\tilde{\mathbf{T}}_y$ by N.

Calibration of unit-level k-nearest neighbour estimates

The proposed calibration of real-valued unit-level kNN predictions builds on the well-known principle that knowledge about the relationship between sample-based (reference) predictor values and the actual predictor values (target) can be exploited and used to generate improved predictions (Brown, 1982; Moody & Woodcock, 1996; Gregoire & Valentine, 1999; Katila et al., 2000).

The proposed calibration aims at lowering extrapolation bias in $\tilde{\mathbf{y}}_i$. Extrapolation bias is more likely to occur when the **X**-values of a target unit lie outside the range(s) covered by the reference units or in a gap not otherwise found in the population distribution of **X** (Stage & Crookston, 2007). It follows from the kNN estimator that target units with similar **X**-values but not covered by the reference units will have similar and potentially seriously biased $\tilde{\mathbf{y}}$-values.

An indicator of the outlier status of a target unit can be obtained by applying the kNN estimator to obtain two kNN estimates of **X** for every target unit. The first, called $\tilde{\mathbf{x}}_i^{ref}$, is estimated exclusively from the reference units (i.e. $\tilde{\mathbf{x}}_i^{ref} = \Sigma_{i \sim k^{\mathrm{j}}} \mathbf{w}_j x_j, j \in \mathbf{s}$), while the second, called $\tilde{\mathbf{x}}_i^{pop}$, is the "true" kNN estimate of **X** as it is estimated from the actual k nearest neighbours in U (i.e. $\tilde{\mathbf{x}}_i^{pop} = \Sigma_{i \sim k^{\mathrm{j}}} \mathbf{w}_j x_j, j \in U$). When $\tilde{\mathbf{x}}_i^{ref} \approx \tilde{\mathbf{x}}_i^{pop}$ and the relationship between **Y** and **X** is approximately multivariate linear in a neighbourhood around $\tilde{\mathbf{x}}_i^{ref}$ and $\tilde{\mathbf{x}}_i^{pop}$, one should expect that $\tilde{\mathbf{y}}_i^{ref} \approx \tilde{\mathbf{y}}_i^{pop}$. Conversely, a large difference between $\tilde{\mathbf{x}}_i^{ref}$ and $\tilde{\mathbf{x}}_i^{pop}$ suggests that $\tilde{\mathbf{y}}_i^{ref} \neq \tilde{\mathbf{y}}_i^{pop}$.

The potential effect of the difference $\tilde{\mathbf{x}}_i^{pop} - \tilde{\mathbf{x}}_i^{ref}$ on a kNN prediction can be approximated if the following simple working multivariate linear model for **Y** is assumed:

$$\mathbf{Y} = F(\mathbf{X})\mathbf{B} + \mathbf{\Xi}, \quad (3)$$

where $\mathbf{Y}$ is an $N \times q$ matrix, $F(\mathbf{X})$ is an $N \times r$ matrix of transforms of $\mathbf{X}$ (detailed in the paragraph starting with "To preserve ...", below), B is an $r \times q$ matrix of least squares regression coefficients ($r > p$) and $\mathbf{\Xi}$ is an $N \times q$ matrix of residuals. Given a sample and a constrained least squares estimate of B, simple linear algebra can be applied to predict, for every unit, a $\hat{\mathbf{y}}_i^{ref}$ from the k-units in $\tilde{\mathbf{x}}_i^{ref}$ and a $\hat{\mathbf{y}}_i^{pop}$ from the k-units in $\tilde{\mathbf{x}}_i^{pop}$. Both are weighted ($\mathbf{w}$) averages of k-individual predictions, using the same weights as in eq. (1). If the model in eq. (3) is correct, the difference $(\hat{\mathbf{y}}_i^{pop} - \hat{\mathbf{y}}_i^{ref})$ becomes an estimate of the extrapolation bias in $\tilde{\mathbf{y}}_i$. Accordingly, the following calibration of unit-level kNN predictions is proposed:

$$\tilde{\mathbf{y}}_i^{cal} = \tilde{\mathbf{y}}_i + (\hat{\mathbf{y}}_i^{pop} - \hat{\mathbf{y}}_i^{ref}). \quad (4)$$

With the calibration in eq. (4), it is possible that $\Sigma_U \hat{y}_i^{pop} - \hat{y}_i^{ref} \neq 0$, which raises the question of whether the calibration should be constrained to sum-to-zero or not. When it can be justified to assume that $\tilde{\mathbf{T}}_y$ is nearly unbiased, a sum-to-zero restriction on the calibration is assured to conserve this desired property. An element-by-element multiplication of $\mathbf{y}_i^{ref}$ in eq. (4) by a set of constants (c_1, ... c_q) accomplishes this. The constants are $c_l = \hat{\mathrm{T}}_{y(l)}^{pop} \times (\hat{\mathrm{T}}_{y(l)}^{ref})^{-1}, l = 1, \ldots, q$, where T stands for a total. In these examples, with small sample sizes, the unconstrained calibration in eq. (4) was used.

To preserve the attraction of kNN as a simple and flexible non-parametric multivariate predictor, the choice of a working model should be straightforward. What is important is that the average change in $\mathbf{Y}$ for a given change in $\mathbf{X}$ is captured adequately by $\mathbf{B}$. To capture all major trends and to minimize model-based extrapolation errors, it is proposed to fit eq. (3) with both $\mathbf{Y}$ and a decorrelated $\mathbf{X}$ scaled to the interval [0,1] and to use a full set of orthogonal constant, linear, quadratic and cubic Bernstein basis functions for the transform functions (F) in eq. (3). There are $D + 1$ Bernstein basis functions of degree D (Lorentz, 1953, p. 30), so one constant, two linear, three quadratic functions and four cubic functions are obtained for each ancillary variable in $\mathbf{X}$. Hence, the dimension of $F(\mathbf{X})$ in eq. (3) is $N \times r$, with $r = 10 \times p$. Decorrelation of $\mathbf{X}$ is done via a Cholesky decomposition (Rencher, 1995, p. 29). The Bernstein $D + 1$ basis functions of degree D are:

$$F(x, d, D) = \binom{D}{d} x^D (1 - x)^{D-d}, \quad d = 0, \ldots, D. \quad (5)$$

Bernstein basis functions are defined on the interval [0,1] with $0 \leq F(x,d,D) \leq 1$. To maintain predictions in the interval [0,1] all regression coefficients were constrained to values between 0 and 1 with a global sum-to-one restriction. The scaling of reference values of $\mathbf{Y}$ should reflect that their ranges are probably less than in the population (Harter, 1970, p. 12). For this study, with strictly positive $\mathbf{Y}$-values, the sample ranges of $\mathbf{Y}$ were expanded by multiplying attribute-specific sample maxima and minima values by factors determined as the average (over 1000 replications) of the ratio of the maxima (minima) in samples of size N and n, drawn at random from a distribution fitted to the observed values of $\mathbf{Y}$ by the method of maximum likelihood. Here, a two-parameter Weibull distribution was chosen as it achieved (trait-by-trait) a good fit to samples of $\mathbf{Y}$. After fitting the model in eq. (3) with the scaled data, the predictions are scaled back to their original scale. Note that the rescaling of $\mathbf{Y}$ is not a requirement; it can be dispensed with. However, the scaling of $\mathbf{Y}$ offers a convenient method for controlling (capping) extrapolations. Without a scaling, examples of (impossible) negative predictions of $\mathbf{Y}$ were seen, which would have to be addressed separately or through a model choice (F) that restricts predictions to their allowed range.

The proposed working model with cubic Bernstein basis functions is flexible and easy to implement, and should catch all major trends in the relationship between $\mathbf{Y}$ and $\mathbf{X}$. If higher order trends are expected, it is easy simply to add higher order Bernstein basis functions. Since the model is not being used for actual prediction purposes, issues of overfitting and collinearity are not a major concern.

Calibration assessment

The proposed calibration procedure is demonstrated in Monte Carlo (MC) simulations of equal probability sampling (without replacement) with either equal-sized clusters ($\mathrm{CLU}_{\mathrm{wor}}$) or single units ($\mathrm{SRS}_{\mathrm{wor}}$) from three artificial populations assembled from actual inventory data. Results are based on 4000 MC replications of a sampling design with k fixed at the value that, for a given sample size, produced the lowest relative root-mean-square error of an estimated total.

On average (across replicated samples) the relationship between $\mathbf{y}_i$ and $\tilde{y}_i$ is ideally linear with a slope (β_1) of one and an intercept (β_0) of zero. An unbiased estimator would meet this condition with no need for a unit-level calibration. However, since unit-level kNN estimates are pulled towards their sample average, and given the aforementioned tendency to generate nearly identical estimates of $\mathbf{Y}$ for target units with an $\mathbf{X}$ outside the range of the

reference units, the desired relationship will typically have a slope greater than one and an intercept less than zero. A successful calibration lowers the slope towards one and raises the intercept towards zero. It does that by extending the range of $\tilde{\mathbf{y}}_i^{cal}$ relative to $\tilde{\mathbf{y}}_i$. To assess the success of the proposed calibration, therefore, sample-based regression coefficients in the linear regression of $\mathbf{y}_i$ on $\tilde{\mathbf{y}}_i^{cal}$ and of $\mathbf{y}_i$ on $\tilde{\mathbf{y}}_i$ are compared. The comparison is extended to coefficients in regressions using the averages $\bar{\tilde{y}}_i^{cal}$ and $\bar{\tilde{y}}_i$ across the MC replications. The statistical significance of departures from the ideal (slope $=1$, intercept $=0$) is assessed with Hotelling's T^2 test statistic (Rencher, 1995, p. 149). The same test is also used to compare the coefficients of the two regressions.

Extending the range of $\tilde{\mathbf{y}}_i$ is also considered a benefit of calibration since the range of $\bar{\tilde{y}}_i$ is clearly less than the natural range of $\mathbf{y}_i$ (Stage & Crookston, 2007). Therefore, the impact of calibration on the natural range of $\tilde{\mathbf{y}}_i^{cal}$ is demonstrated. If a calibration achieves the positive shift in the regression, one should also expect to see a reduction in estimation errors, but not necessarily a decline in the bias of a total. Since the total bias in $\tilde{\mathbf{y}}_i$ and $\tilde{\mathbf{y}}_i^{cal}$ is expected to be approximately equal (it depends foremost on k and the chosen weighting scheme), the ratio of bias-adjusted residuals is taken as an estimator of the relative reduction in prediction errors.

$$R_{error}^{cal} = (\tilde{\mathbf{y}}_i^{cal} - \mathbf{y}_i - Mean(\tilde{\mathbf{y}}_i^{cal} - \mathbf{y}_i)) \times (\tilde{\mathbf{y}}_i^{cal} - \mathbf{y}_i - Mean(\tilde{\mathbf{y}}_i - \mathbf{y}_i))^{-1}. \quad (6)$$

Applications

The proposed model-assisted procedure for reducing extrapolation bias in kNN applications is demonstrated with three examples using actual forest inventory data from Italy (IT), and from two small artificial forests compiled from actual inventory data from Finland and the USA. The data from Finland and the USA have been used in an earlier study on model-based estimation of root mean-squared errors in kNN applications (Magnussen et al., 2009). The Finnish data were collected from forests on mineral and peat soils. These data have been combined into a single data set called MIN&PEAT. The US data were originally from two disjoint but similar areas (called FIA1 and FIA2 in Magnussen et al. (2009). Here, they are combined under the name of F1&2.

IT. This population of $N=312$ forest compartments is located in the forests of Trentino-Alto Adige (northern Italy), as detailed in Baffetta et al. (2009). The interest variable ($\mathbf{Y}$) is timber volume (VOL, $m^3\,ha^{-1}$). Stand-level volume estimates were obtained by calipering each tree for diameter at breast height and applying local volume tables. The mean stand volume was 318 $m^3\,ha^{-1}$ with an among-stand standard deviation of 128 $m^3\,ha^{-1}$. The ancillary variables ($X1$, $X2$, ….., $X6$) are the average within-stand digital numbers of spectral bands 1, 2, 3, 4, 5 and 7 from concurrent Landsat 7 ETM+ imagery. The Euclidean distance was used to determine the neighbour structure in $\mathbf{X}$-space. Figure 1 shows scatterplots of stand volume against the six ancillary variables after scaling $\mathbf{X}$ to the interval [0,1]. The ancillary variables are all negatively correlated with VOL (-0.58 to -0.37). Calibration results are for a simple random sample size of $n=20$ and $k=6$.

MIN&PEAT. This is an artificial population of $N=3912$ units arranged in 1304 clusters each with 3 units in a 1×3 array configuration. Units are on a grid with 163 rows and 24 columns. A unit is a quarter of a Landsat 7 ETM+ image pixel. Each unit has a forest inventory plot providing $\mathbf{y}_i$, $i=1, \ldots, N$. Unit-level data came from forest inventory plots in the 9th Finnish National Forest Inventory located in North Karelia and South Savo. The variables of interest are $Y1$ = quadratic mean diameter (cm) of tree stems 1.3 m above ground (QMD), $Y2$ = basal area ($m^2\,ha^{-1}$) of tree stems 1.3 m above ground (BA), and $Y3$ = total stem volume in ($m^3\,ha^{-1}$) (VOL). A trivariate single-index-model transform

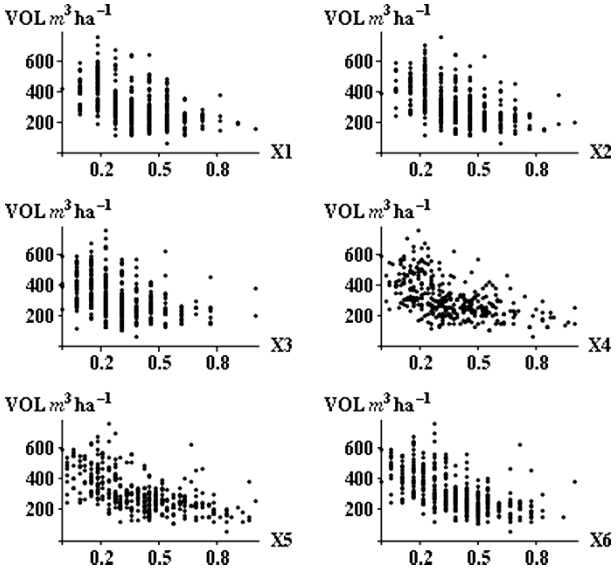

Figure 1. Scatterplots of stand stem volume (VOL $m^3\,ha^{-1}$) against six scaled [0,1] and decorrelated ancillary variables ($X1$, …, $X6$). Site = IT.

(Härdle et al., 1993) of a nine-dimensional vector ($\mathbf{x}_i$) of Landsat 7 ETM+ pixel data was used as ancillary variables *X*1, *X*2 and *X*3. The ancillary variables were scaled to the interval [0,1]. Further details are in Tomppo and Halme (2004). Figure 2 shows scatterplots of VOL against the three ancillary variables. The spatial trend in **X**-values was modelled as $|\sin(\text{row} \times 24^{-1}) \times \cos(\text{row} \times 163^{-1})|$ and units were placed on the grid to minimize the squared departures from this trend. This placement generated an intracluster correlation (Cochran, 1977, p. 209) of 0.07 (QMD), 0.09 (BA) and 0.10 (VOL). Figure 3 shows maps of the spatial distribution of *X*3 and VOL. Data averages and standard deviations (in parentheses) are: QMD 19.2 ($\pm$ 7.0) cm, BA 20.1 ($\pm$ 7.9) $m^2\,ha^{-1}$, and VOL 142.4 ($\pm$ 90.7) $m^3\,ha^{-1}$. All results are based on cluster sampling

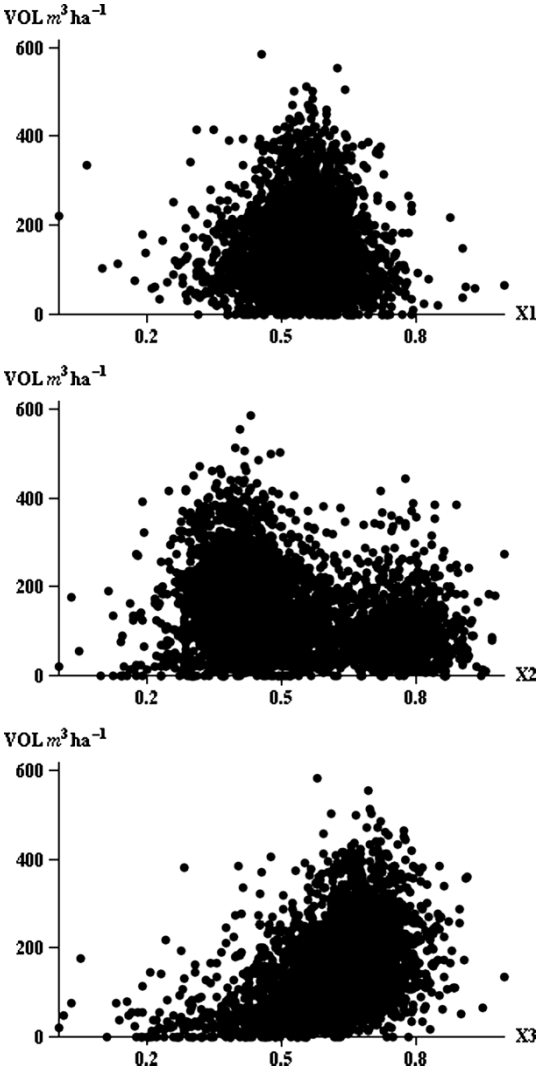

Figure 2. Scatterplots of plot stem volume (VOL $m^3\,ha^{-1}$) against three scaled [0,1] and decorrelated ancillary variables (*X*1, *X*2, *X*3). Site = MIN&PEAT.

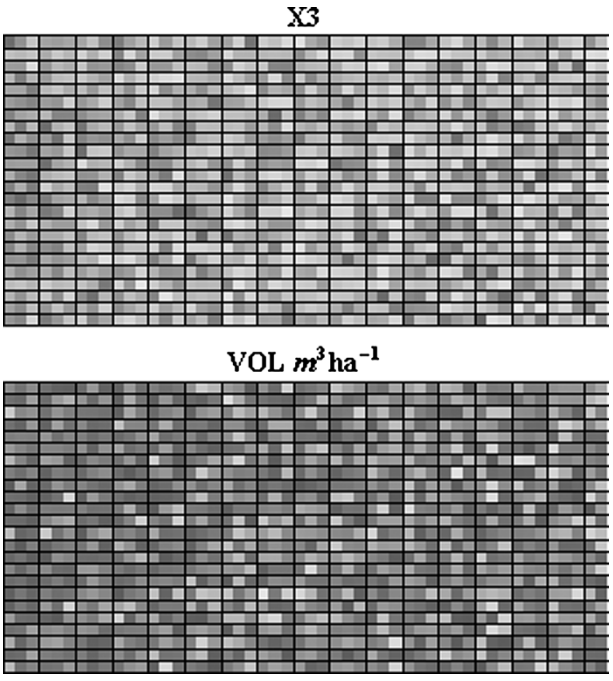

Figure 3. Spatial distribution of *X*3 (top) and VOL (bottom) in MIN&PEAT. (Only the first 50 rows are shown.) Clusters (sampling unit) of three units arranged in a 1×3 array are indicated with black gridlines.

with $n = 20$ and $k = 6$. Multivariate estimators are used throughout, but only VOL results are reported.

FIA1&2. This is an artificial population of $N = 4260$ units arranged in 1065 clusters each with 4 units in a 2×2 array configuration. Units are on a grid with 142 rows and 30 columns. Unit-level data came from the Forest Inventory and Analysis (FIA) program of the US Forest Service (Bechtold & Patterson, 2005) and represent forested areas in Minnesota. A unit is a Landsat 7 ETM+ pixel with a colocated FIA subplot providing $\mathbf{y}_i$, $i = 1, \ldots, N$. Variables in $\mathbf{y}$ are: *Y*1 = number of trees ha^{-1} (TPH), *Y*2 = basal area ($m^2\,ha^{-1}$) (BA), and *Y*3 = merchantable volume ($m^3\,ha^{-1}$) (VOL). A trivariate single-index-model transform (Härdle et al., 1993) of a 12-dimensional vector ($\mathbf{x}_i$) of Landsat 7 ETM+ derived pixel data (Magnussen et al., 2009) was used as ancillary variables (*X*1, *X*2 and *X*3). The ancillary variables were scaled to the interval [0,1]. Figure 4 shows scatterplots of VOL against the three ancillary variables. The spatial trend in **X**-values was modelled as $|5\ \sin(\text{column} \times 142^{-1}) \times \cos(0.5\ \text{row} \times 30^{-1})|$ and units were placed on the grid to minimize the squared departures from this trend. This placement procedure generated an intracluster correlation of 0.08 (TPH), 0.07 (BA) and 0.06 (VOL). Figure 5 shows maps of the spatial distribution of *X*3 and VOL. Data averages and standard

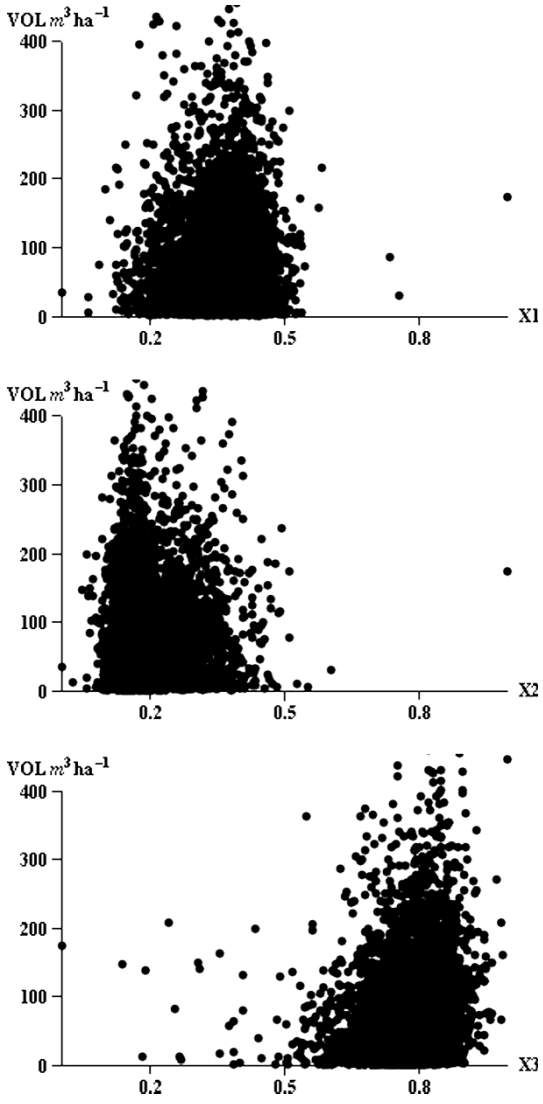

Figure 4. Spatial distribution of *X*3 (top) and VOL (bottom) in MIN&PEAT. (Only the first 50 rows are shown.) Clusters (sampling unit) of three units arranged in a 1×3 array are indicated with black gridlines.

deviations (in parentheses) are: TPH 519 ($\pm$ 420) ha^{-1}, BA 15.7 ($\pm$ 11.9) m^2 ha^{-1}, and VOL 86.3 ($\pm$78.6) m^3 ha^{-1}. All results are based on cluster sampling with $n = 20$ and $k = 8$ and multivariate estimators, but only VOL results are reported.

Results

IT. In a random sample of size 20 taken from a population of size 312 the range of the six predictors was reduced by an average between 30% (*X*2) and 37% (*X*5), and the median gap-size between neighbouring **X**-values was approximately 12 times larger than in the population. The average distance between two **X**-values in a sample (reference units) was 0.39 versus 0.18 in the population. A reduced range

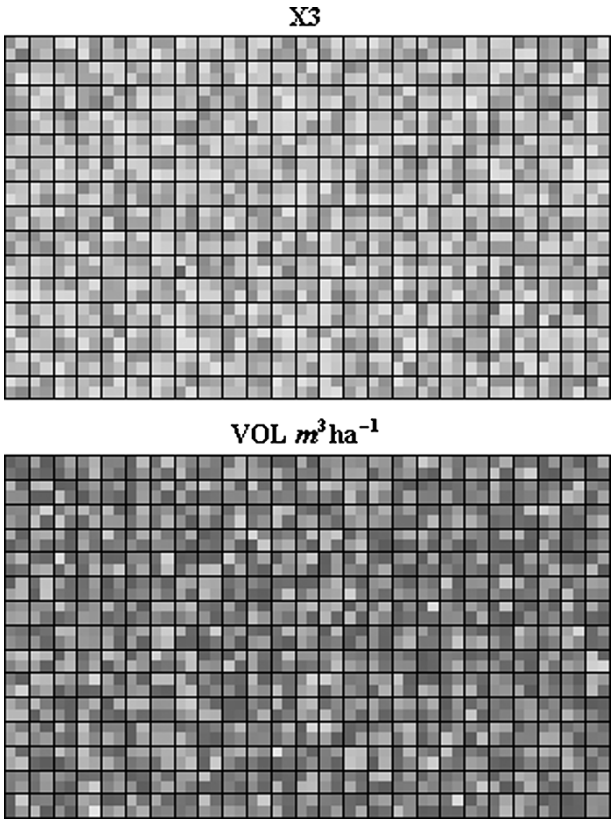

Figure 5. Spatial distribution of *X*3 (top) and VOL (bottom) in FIA1&2. (Only the first 50 rows are shown.) Clusters (sampling unit) of four units arranged in a 2×2 array are indicated with black gridlines.

and larger gaps in **X**-values of the reference units will both contribute to extrapolation bias. For a sample size $n = 20$ and $k = 6$ the non-constrained calibration lowered the MC estimate of bias of total volume from 3.3% to 1.6%. The kNN calibration achieved, in 92% of the MC replications, a shift in the regression of the true y on the kNN prediction towards the desired 1:1 line. For calibrated kNN predictions, the average sample-based estimate of the slope was 1.18 ± 0.04 compared with 1.46 ± 0.08 for non-calibrated kNN predictions. Corresponding estimates of intercepts were -60.2 ± 11.7 and -155.9 ± 23.83, respectively. Hence, calibration reduced the deviations in slope and intercept from the desired values of 1 and 0 by approximately 60%. Figure 6 shows the 95% bivariate quantile envelope for the estimated slopes and intercepts. It is clear that calibration, as a rule, not only shifted the regressions towards the 1:1 line but also lowered (by approximately 50%) the variation of the estimates. Both ellipsoids in Figure 6 include the locus of a zero intercept and slope of one. Only 1% of the calibrated regressions differed significantly at the 5% level from a 1:1 line, whereas 10% of the non-calibrated regressions did. When the calibration

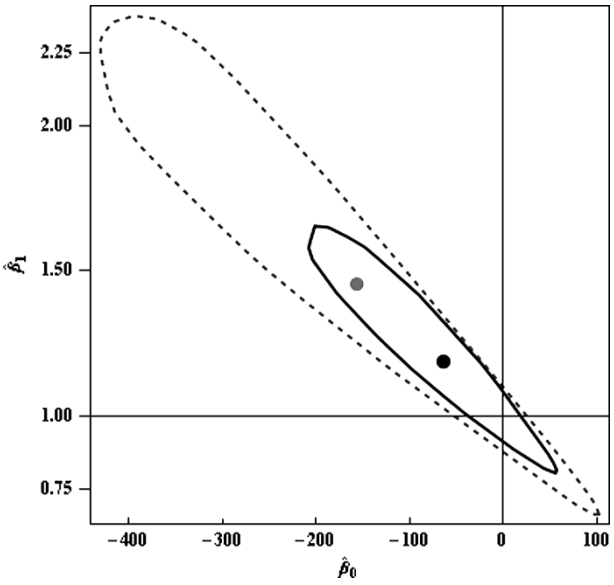

Figure 6. Bivariate quantile (95%) envelope for intercepts (β_0) and slopes (β_1) in the regression of **Y**-values of VOL on *k*-nearest neighbour (kNN) predictions. Dashed line = non-calibrated kNN predictions; solid line = calibrated kNN predictions. Averages of sample-based estimates of β_0 and β_1 are indicated (grey = non-calibrated; black = calibrated). Site = IT.

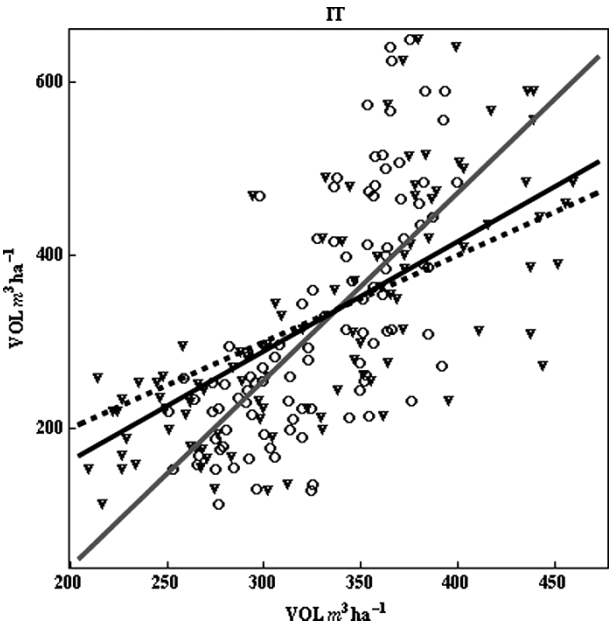

Figure 7. Scatterplot of 100 random selections of unit-level values of volume (VOL $m^3\,ha^{-1}$) against the average *k*-nearest neighbour prediction. Circles = non-calibrated; triangles = calibrated. Ordinary least squares trend lines are indicated (grey = non-calibrated; black = calibrated). A 1:1 line (black, dashed) is provided for reference. Site = IT.

occasionally introduced a negative shift in the regression, it was always minor and of no practical consequence.

In Figure 7, **Y** is plotted against the average of 4000 unit-level kNN predictions. The trend line for non-calibrated predictions had a slope of 2.16 ± 0.14 and an intercept of -393.1 ± 47.8. In comparison, the slope and intercept of calibrated predictions were 1.26 ± 0.1 and -88.0 ± 27.7, respectively. Both trend lines deviated significantly from a 1:1 line ($p < 0.01$). The difference between sample-based and average-population-based parameter estimates is due to attenuation of sample-based estimates owing to the sampling errors in individual kNN predictions (Fuller, 1987, p. 3).

Bias-adjusted errors of calibrated kNN predictions were $20 \pm 8\%$ lower than for non-calibrated predictions. The chance of an inflated error was 0.2%. Calibrated predictions displayed a wider range of values than non-calibrated predictions. Estimated stand-level volume per hectare varied from 60.6 $m^3\,ha^{-1}$ to 760 $m^3\,ha^{-1}$, while calibrated predictions varied (on average) from 152.6 $m^3\,ha^{-1}$ to 513.1 $m^3\,ha^{-1}$ or 52% of the actual range. In contrast, non-calibrated predictions were (on average) between 215.0 $m^3\,ha^{-1}$ and 443.8 $m^3\,ha^{-1}$ or 33% of the actual range. The correlation between y and $\tilde{y}_i^{cal}$ was, on average, 15.4% larger than between y and $\tilde{y}_i$. A stronger correlation was seen in 92% of the MC replications. Calibration also produced a 7.6% lower mean absolute difference (MAD) between a kNN prediction and its true value.

MIN&PEAT. The average range of **X**-values in a 0.5% sample was only 33% ($X1$), 57% ($X2$) and 42% ($X3$) of the full range in the population. Gaps in the **X**-values of reference units were, on average, about 100 times larger than in the population. The average distance of two reference units was approximately eight times larger than in the population. From these figures one perceives a non-trivial risk of extrapolation bias. For a sample size of 20 clusters (i.e. 60 units) and $k = 6$ the MC estimate of bias of total volume was 1.2% for both non-calibrated and unconstrained calibrated kNN predictions. Calibration achieved a shift in the regression of the true y on the kNN prediction towards a 1:1 line in 88% of the MC replications. Sample-based regressions with calibrated predictions had an average slope and intercept of 1.03 ± 0.11 and -4.44 ± 14.44, respectively. Corresponding estimates for the non-calibrated predictions were 1.12 ± 0.14 and -17.54 ± 18.11. Thus, the calibration reduced by 10% the departure of the slope from 1 and by approximately 75% the departure of the intercept from 0. Figure 8 shows the 95% bivariate quantile envelope for slopes and intercepts. Calibration has shifted the regressions towards the 1:1 line and lowered the variation in estimated slopes and intercepts by about 30%.

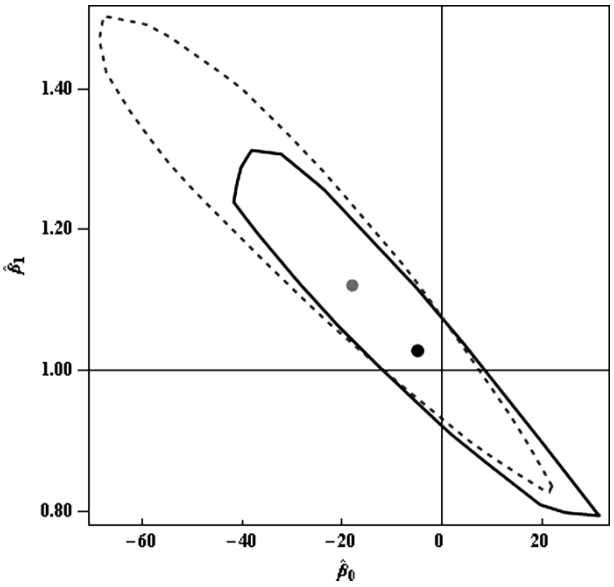

Figure 8. Bivariate quantile (95%) envelope for intercepts (β_0) and slopes (β_1) in the regression of **Y**-values of VOL on k-nearest neighbour predictions. Dashed line = non-calibrated predictions; solid line = calibrated predictions. Averages of sample-based estimates of β_0 and β_1 are indicated (grey = non-calibrated; black = calibrated). Site = MIN&PEAT.

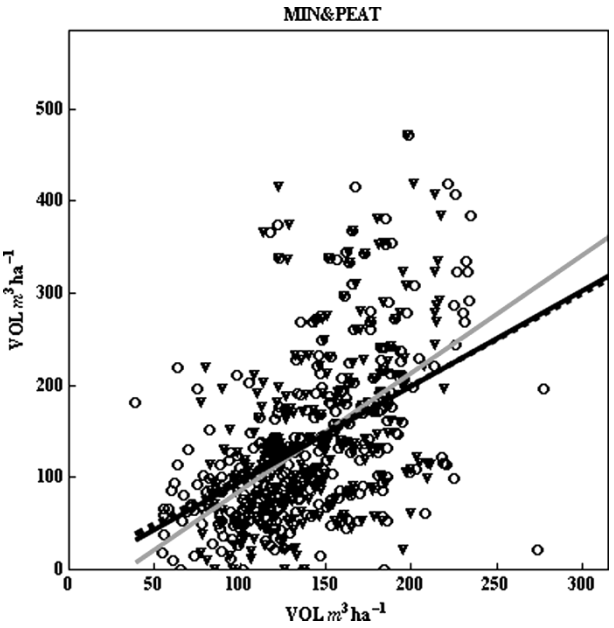

Figure 9. Scatterplot of 100 random selections of unit-level values of volume (VOL $m^3\,ha^{-1}$) against the average k-nearest neighbour prediction. Circles = non-calibrated; triangles = calibrated. Ordinary least squares trend lines are indicated (grey = non-calibrated; black = calibrated). A 1:1 line (black, dashed) is provided for reference. Site = MIN&PEAT.

Less than 1% of the regressions deviated significantly at the 5% level from a 1:1 line (Hotelling's T^2 test).

When y was regressed on the average kNN prediction the trend line for non-calibrated predictions had a slope of 1.27 ± 0.03 and an intercept of -41.12 ± 5.03 (Figure 9), whereas in regression with averages of calibrated predictions the slope was 1.05 ± 0.03 and the intercept -8.78 ± 4.17. Only the latter was not significantly different from a 1:1 line ($p = 0.08$).

Bias-adjusted errors of calibrated kNN predictions were just $2.7 \pm 1.8\%$ lower than for non-calibrated predictions. The chance that calibration introduces a small inflation of bias-adjusted errors was 5.8%. Calibrated predictions of unit volume per hectare varied (on average) from 14.1 $m^3\,ha^{-1}$ to 359.0 $m^3\,ha^{-1}$. Corresponding numbers for the non-calibrated predictions are 44.4 $m^3\,ha^{-1}$ and 271.7 $m^3\,ha^{-1}$. As a result, the calibrated predictions capture approximately 60% of the actual range of 0 − 585 $m^3\,ha^{-1}$, while non-calibrated predictions only capture approximately 40% of the actual range. The correlation between y and $\tilde{y}_i^{cal}$ was, on average, 6.5% larger than the correlation between y and $\tilde{y}_i$. Calibration lowered MAD by approximately 1.0%.

FIA1&2. As in the previous example, a much smaller portion (70 − 73%) of the range of **X**-values was seen in the reference units compared with the full range of values in the population. In addition, there were much larger gaps (on average, approximately 80 times larger) and a larger (130%) average distance between reference **X**-values than seen in the population. From these figures the opportunity to reduce extrapolation bias appears favourable. With a sample size of 20 clusters (i.e. 80 units) and $k = 8$ the MC estimate of bias of total volume was 0.2% for the calibrated and 1.1% for the non-calibrated kNN predictions. Calibration shifted the regression of the true y on the kNN prediction towards a 1:1 line in 94% of the MC replications. Sample-based regressions with calibrated predictions had an average slope and intercept of 1.05 ± 0.20 and -3.72 ± 15.90, respectively. Corresponding estimates for the non-calibrated predictions were 1.09 ± 0.22 and -7.37 ± 18.52. Figure 10 illustrates the modest effects of calibration in the form of slightly shifted 95% bivariate quantile envelopes for slopes and intercepts.

When y was regressed on the average calibrated kNN prediction the trend line had a slope of 1.10 ± 0.06 and an intercept of -9.07 ± 5.62 (Figure 11). Departures from a 1:1 line were not significant ($p = 0.27$). In contrast, the trend line for the average non-calibrated predictions had a slope 1.30 ± 0.08 and an intercept of -26.4 ± 6.83 which constitute a significant departure from a 1:1 line ($p = 0.001$).

Unit-level volume per hectare had a data range of 300 $m^3\,ha^{-1}$ (= 100%). Calibrated predictions

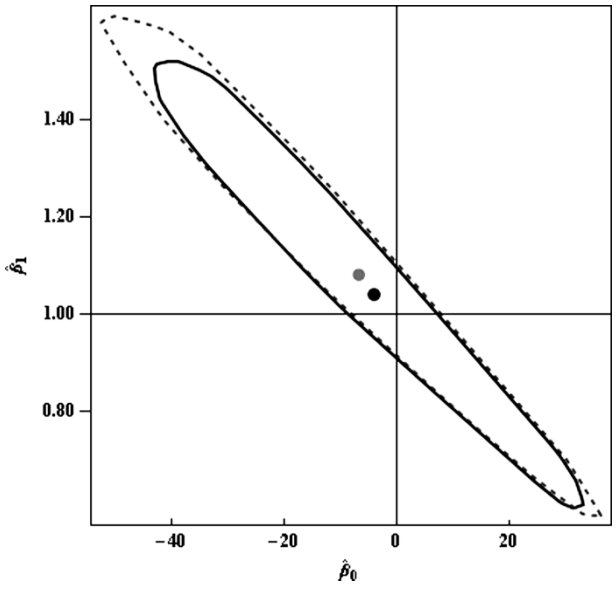

Figure 10. Bivariate quantile (95%) envelope for intercepts (β_0) and slopes (β_1) in the regression of **Y**-values of VOL on *k*-nearest neighbour predictions. Dashed = non-calibrated predictions; solid line = calibrated predictions. Averages of sample based estimates of β_0 and β_1 are indicated (grey = non-calibrated; black = calibrated). Site = FIA1&2.

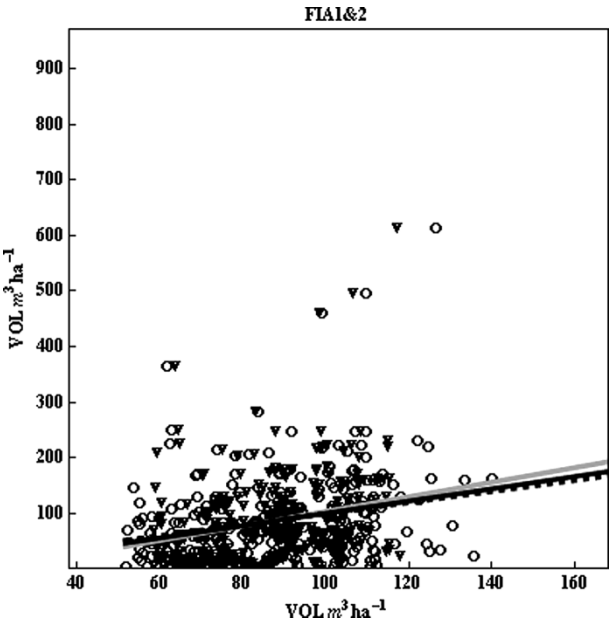

Figure 11. Scatterplot of 100 random selections of unit-level values of volume (VOL $m^3\,ha^{-1}$) against the average *k*-nearest neighbour prediction. Circles = non-calibrated; triangles = calibrated. Ordinary least squares trend lines are indicated (grey = non-calibrated; black = calibrated). A 1:1 line (black, dashed) is provided for reference. Site = FIA1&2.

displayed an average range of 212 $m^3\,ha^{-1}$ (71%), whereas non-calibrated had a range of just 149 $m^3\,ha^{-1}$ (50%). All other calibration effects were negligible.

Discussion

Despite its popularity, it is well known that unit-level kNN predictions can be seriously biased (LeMay et al., 2008; Eskelson et al., 2009; McRoberts, 2009). Out-of-sample extrapolation is a major source of bias regardless of whether extrapolation goes beyond the range of the sample data (**X**) or into gaps of **X** that are artefacts of sampling (Stage & Crookston, 2007). The calibration technique proposed here mitigates both types of bias, but not the bias that stems from a large variation in **Y**-values for a given **X**. The MIN&PEAT and F1&2 examples have a large variation in **Y** for a given **X** which explains, in part, why the attempt to reduce extrapolation bias was (seemingly) less efficient in these two examples than in the example from Italy. The much smaller sample fractions in the case of MIN&PEAT and F1&2 create a greater risk of extrapolation bias, but if the relationship between **X** and **Y** is rather weak the bias generated by having widely different **Y**-values for a single **X**-value may be the most important source of bias (Stage & Crookston, 2007).

When the **X**-variables used in a kNN application contain significant information about the attributes of interest (**Y**), a prediction based on the *k* nearest reference units in a sample will be different from a prediction based on the actual *k* nearest units in the population. The latter would be a better prediction, in terms of both bias and precision. The challenge is to quantify the expected shift between the two predictions and then adjust sample-based predictions accordingly. Calibration is therefore invariably model based (Brown, 1982). In applications with sample sizes *n* much smaller than the population size *N*, the *k* reference units selected for making a kNN prediction will in most cases be different from the actual *k* nearest neighbours in the populations. The effect of this selection differential depends on the nature and strength of the relationship between **X** and **Y**. The examples presented here suggest that the most important benefit of calibration is a sizeable extension of the range of predicted values towards their actual range. This is accomplished without adverse impact on overall bias, with a high chance of modest positive effects on errors, and with the assurance that in the long run the trend line between actual and predicted values will be closer to a 1:1 line. Results from FIA1&2 suggest that these benefits materialize even when the relationship between **X** and **Y** is weak. The scaling of the **Y**-values pursued here to control extrapolation errors has not contributed to the partial recovery of the natural range of the **Y**-variables. All operators acting on **Y** are strictly linear in **Y**; hence, the effect of a scaling is completely removed by an inverse scaling.

The proposed calibration does not seem to have any adverse effects on the correlation among predicted **Y**-values. In the three examples, the pairwise correlation coefficients of calibrated kNN predictions were never larger than the coefficients of non-calibrated predictions. On average, they were 0.03 closer to the actual population values.

Calibration should be straightforward and easy to implement, otherwise the kNN technique would lose its appeal (Haara et al., 1997). The present modelling approach (with decorrelated **X** variables, **X** and **Y** confined to the interval [0,1], and orthogonal Bernstein basis functions) is easy to implement in any programming language. The Bernstein functions were limited to capture constant, linear and cubic trends, but when deemed appropriate an extension to quartic or even quintic trends is easy to implement. Alternatively, a model-assisted calibration is also possible with a strictly non-linear model as long as the chosen model can be supported by external knowledge.

A decorrelation of **X**-values is routine unless there is redundancy in the selected **X**-variables, in which case one or more redundant variables should be eliminated (Li & Wang, 2007; Wang & Xia, 2008). The orthogonal basis functions effectively remove collinearity issues and, finally, confining **Y** to the unit intervals offers an effective control on extrapolations. Prior knowledge of the natural upper and lower limits of **Y** is preferable to simulation-based or parametric estimation of these limits (Sarhan & Greenberg, 1962). In large populations, the time to find the k nearest neighbours to each unit does add to computing time, but efficient search algorithms are available (Finley & McRoberts, 2008).

Given that calibrated predictions, just like the original kNN predictions, can be viewed as (model-based) proxies for the actual **Y**-values, the analyst can still use the empirical (probability-based) difference estimator (Baffetta et al., 2009) or any other suitable estimator (Magnussen, McRoberts & Tomppo, 2009, unpublished results) for calculating the sampling variance of an estimated total. An issue may arise on how to count degrees of freedom (Särndal et al., 1992, p. 222). When the spatial consistency of kNN predictions is important a correction procedure by Barth et al. (2009) may be applied to calibrated predictions.

In conclusion, calibration of kNN predictions may be recommended as a routine when unit-level predictions are desired in their own right (Maselli et al., 2005; Eskelson et al., 2009), and for small-area estimation (Katila, 2006). In small-area estimation, the net effect of using sample-based k-nearest units instead of the actual k-nearest units can be expected to vary among subpopulations owing to varying degrees of extrapolation (Tomppo et al., 1999; Fehrmann et al., 2008). A preliminary assessment of the need for a calibration (McRoberts, 2009) may take more time than implementing the proposed calibration procedure.

Acknowledgements

Data for the IT population were kindly made available by Dr Piermaria Corona, University of Tuscany, Department of Forest Environment and Resources. We are grateful to three anonymous journal referees and the Editor for numerous constructive and helpful comments and suggestions to an earlier version of this manuscript

References

Alt, H. (2001). *The nearest neighbor. Computational Discrete Mathematics: Advanced Lectures*, *2122*, 13–24.

Baffetta, F., Fattorini, L., Franceschii, S. & Corona, P. (2009). Design-based approach to the kNN technique for coupling field and remotely sensed data in forest surveys. *Remote Sensing of Environment*, *113*, 463–475.

Barth, A., Wallerman, J. & Ståhl, G. (2009). Spatially consistent nearest neighbor imputation of forest stand data. *Remote Sensing of Environment*, *113*, 546–553.

Bechtold, W. A. & Patterson, P. L. (2005). *The enhanced forest inventory and analysis program—National sampling design and estimation procedures*. USDA Forest Service, General Technical Report, SRS–80.

Brown, P. J. (1982). Multivariate calibration. *Journal of the Royal Statistical Society, Series B*, *44*, 287–321.

Cochran, W. G. (1977). *Sampling techniques*. New York: Wiley.

Eskelson, B. N. I., Temesgen, H., Lemay, V., Barrett, T. M., Crookston, N. L. & Hudak, A. T. (2009). The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases. *Scandinavian Journal of Forest Research*, *24*, 235–246.

Fehrmann, L., Lehtonen, A., Kleinn, C. & Tomppo, E. (2008). Comparison of linear and mixed-effect regression models and a k-nearest neighbour approach for estimation of single-tree biomass. *Canadian Journal of Forest Research*, *38*, 1–9.

Finley, A. O. & McRoberts, R. E. (2008). Efficient k-nearest neighbor searches for multi-source forest attribute mapping. *Remote Sensing of Environment*, *112*, 2203–2211.

Franco-Lopez, H., Ek, A. R. & Bauer, M. E. (2001). Estimation and mapping of forest stand density, volume, and cover type using the k-nearest neighbors method. *Remote Sensing of Environment*, *77*, 251–274.

Fuller, W. A. (1987). *Measurement error models*. New York: Wiley.

Gregoire, T. G. & Valentine, H. T. (1999). Composite and calibration estimation following 3P sampling. *Forest Science*, *45*, 179–185.

Haara, A., Maltamo, M. & Tokola, T. (1997). The k-nearest-neighbour method for estimating basal area diameter distribution. *Scandinavian Journal of Forest Research*, *12*, 200–208.

Hansen, M. H., Madow, W. G. & Tepping, B. J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, *78*, 776–793.

Härdle, W., Hall, P. & Ichimura, H. (1993). Optimal smoothing in single-index models. *Annals of Statistics*, *21*, 157–178.

Harter, H. L. (1970). *Order statistics and their use in testing and estimation* (Vol. II, p. 172). Ohio: Aerospace Research Laboratories.

Katila, M. (2006). Empirical errors of small area estimates from the multisource national forest inventory in eastern Finland. *Silva Fennica*, *40*, 729–742.

Katila, M., Heikkinen, J. & Tomppo, E. (2000). Calibration of small-area estimates for map errors in multisource forest inventory. *Canadian Journal of Forest Research*, *30*, 1329–1339.

Krause, A. & Pinheiro, J. (2007). Modeling and simulation to adjust values in presence of a regression to the mean effect. *American Statistician*, *302–307*, 61,.

LeMay, V., Maedel, J. & Coops, N. C. (2008). Estimating stand structural details using nearest neighbor analyses to link ground data, forest cover maps, and Landsat imagery. *Remote Sensing of Environment*, *112*, 2578–2591.

Li, B. & Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association*, *102*, 997–1008.

Lorentz, G. G. (1953). *Bernstein polynomials* (2nd ed). Toronto: University of Toronto Press.

McRoberts, R. E. (2009). Diagnostic tools for nearest neighbors techniques when used with satellite imagery. *Remote Sensing of Environment*, *113*, 489–499.

Magnussen, S., McRoberts, R. E. & Tomppo, E. (2009). Model-based mean square error estimators for k-nearest neighbour predictions and applications using remotely sensed data for forest inventories. *Remote Sensing of Environment*, *113*, 476–488.

Maselli, F., Chirici, G., Bottai, L., Corona, P. & Marchetti, M. (2005). Estimation of Mediterranean forest attributes by the application of k-NN procedures to multitemporal Landsat ETM plus images. *International Journal of Remote Sensing*, *26*, 3781–3796.

Meng, Q. M., Cieszewski, C. J., Madden, M. & Borders, B. E. (2007). *k* nearest neighbor method for forest inventory using remote sensing data. *GIScience & Remote Sensing*, *44*, 149–165.

Moody, A. & Woodcock, C. E. (1996). Calibration-based models for correction of area estimates derived from coarse resolution land-cover data. *Remote Sensing of Environment*, *58*, 225–241.

Rencher, A. C. (1995). *Methods of multivariate analysis*. New York: Wiley.

Sarhan, A. E. & Greenberg, B. G. (1962). *Contributions to order statistics*. New York: Wiley.

Särndal, C. E., Swensson, B. & Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer.

Scott, D. W. (1992). *Multivariate density estimation: Theory, practice and visualization*. New York: Wiley.

Stage, A. R. & Crookston, N. L. (2007). Partitioning error components for accuracy-assessment of near-neighbour methods of imputation. *Forest Science*, *53*, 62–72.

Tomppo, E. & Halme, M. (2004). Using coarse scale forest variables as ancillary information and weighting of variables in k-N-N estimation: a genetic algorithm approach. *Remote Sensing of Environment*, *92*, 1–20.

Tomppo, E., Goulding, C. & Katila, M. (1999). Adapting Finnish multi-source forest inventory techniques to the New Zealand preharvest inventory. *Scandinavian Journal of Forest Research*, *14*, 182–192.

Tomppo, E., Olsson, H., Stahl, G., Nilsson, M., Hagner, O. & Katila, M. (2008). Combining national forest inventory field plots and remote sensing data for forest databases. *Remote Sensing of Environment*, *112*, 1982–1999.

Wang, H. & Xia, Y. (2008). Sliced regression for dimension reduction. *Journal of the American Statistical Association*, *103*, 811–821.